SEMIOSEM et PROXIMA : mesures sémiotiques de similarité et de proximité conceptuelles

Xavier Aimé[1], Frédéric Fürst[2], Pascale Kuntz[1], Francky Trichet[1]

[1] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
Université de Nantes, équipe COD - Connaissances & Décision
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
xavier.aime@paris-sorbonne.fr
{pascale.kuntz,francky.trichet}@univ-nantes.fr

[2] MIS - Modélisation, Information et Systèmes
Université de Picardie - Jules Verne
33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr

Abstract :

Cet article présente deux nouvelles mesures de similarité et proximité entre concepts d'une ontologie, qui ont pour originalité, d'une part, de bien distinguer ces deux notions et, d'autre part, de tirer parti des trois composantes sémiotiques des concepts : intension, extension et expression. Les calculs de proximité et de similarité reposent donc sur les propriétés des concepts, sur leurs instances et sur leurs termes. Des ressources additionnelles à l'ontologie sont également utilisées, en particulier un corpus de textes. La mesure SEMIOSEM évalue la similarité entre deux concepts, c'est-à-dire leur analogie, en s'appuyant sur la comparaison entre leurs prototypes et sur le fait qu'ils partagent des instances communes et des termes communs. La mesure PROXIMA évalue la proximité entre deux concepts, c'est-à-dire le fait qu'ils soient cognitivement liés, en mesurant la densité des propriétés qui les relient, et l'apparition conjointe dans un corpus de textes des termes désignant ces concepts ou leurs instances. Nous présentons également les premiers tests réalisés qui permettent de comparer nos mesures avec les mesures de similarité existantes et de comparer SEMIOSEM et PROXIMA entre elles.
Mots-clés : Mesure sémantique, Similarité, Proximité, Sémiotique.

1. Introduction

La notion de *proximité sémantique* est aussi vieille que celle de réseau sémantique et remonte aux travaux de Quillian [Quillian (1968)]. La notion de proximité introduite par Quillian correspond à l'association d'idée et mesure

donc à quel point une notion peut venir à l'esprit lorsque une autre notion est évoquée. Mais cette notion générale de proximité peut recouvrir au moins deux types d'association : la *similarité conceptuelle*, c'est-à-dire une analogie entre deux concepts qui partagent un certain nombres de caractéristiques et la *proximité conceptuelle* au sens strict, c'est-à-dire un lien cognitif établi entre deux concepts dans un certain contexte. Par exemple, un article de recherche et un relecteur sont des concepts proches, car évoqués ensemble dans un même contexte, mais ils ne sont pas du tout similaires. Par contre, un article de recherche et une profession de foi politique sont très similaires, dans leur nature, voire dans leur structure, mais ne sont généralement pas proches.

Il est à noter que cette distinction entre proximité et similarité a été introduite depuis longtemps en psychologie cognitive dans le cadre de la théorie de la Gestalt [Koffka (1935)]. Le gestaltisme théorise la perception au travers de différentes lois qui précisent la façon dont les formes globales sont appréhendées à partir des détails perçus par l'œil. La loi de proximité stipule que nous regroupons d'abord les éléments proches et la loi de similitude que les détails distants sont regroupés selon les similarités qu'ils présentent.

En ingénierie des connaissances, le développement des ontologies a conduit à la définition de plusieurs mesures de similarité entre concepts, mesures qui sont particulièrement utiles pour l'alignement d'ontologies [Euzenat & Valtchev (2004)], pour l'analyse de texte [Budanitsky & Hirst (2006)] ou la recherche d'information [Hliaoutakis *et al.* (2006)]. Deux approches sont principalement utilisées : celles qui reposent sur l'utilisation des liens hiérarchiques entre concepts et celles qui utilisent des ressources externes (en particulier des textes). Nous mêmes avions proposé une première version de la mesure de similarité SEMIOSEM, qui exploitait un corpus de textes [Aimé *et al.* (2009b)]. Cependant, ces mesures - dont SEMIOSEM dans sa première version - ne permettent généralement pas de bien distinguer entre proximité au sens strict et similarité. Par exemple, deux concepts évoqués souvent dans les mêmes textes sont sans doute proches, mais pas forcément similaires, c'est le cas de l'article de recherche et du relecteur. Deux concepts tels qu'article de recherche et profession de foi politique peuvent être des concepts frères dans une ontologie, et sont très similaires sans pour autant être proches cognitivement parlant.

Partant du constat que similarité et proximité sont deux notions proches mais non similaires, nous proposons dans ce papier deux mesures sémantiques distinctes : une nouvelle version de la mesure de similarité SEMIOSEM et une mesure de proximité PROXIMA. La première évalue l'analogie entre deux

concepts, et est donc fondée uniquement sur les caractéristiques des concepts (propriétés, termes, instances) indépendamment de la structure de l'ontologie qui les contient. SEMIOSEM n'utilise pas non plus de ressources additionnelles, comme des textes. PROXIMA évalue la proximité entre deux concepts, au sens défini précédemment, et, outre l'ontologie contenant les deux concepts, utilise un corpus de textes représentatifs du domaine.

La suite de cet article est structurée comme suit. La section 2 présente le modèle formel au travers duquel nous manipulons les ontologies. Les sections 3 et 4 décrivent respectivement SEMIOSEM et PROXIMA. La section 5 présente des résultats expérimentaux.

2. Un modèle d'ontologie sémiotique

Le modèle ontologique que nous utilisons pour définir nos mesures incorpore des éléments non standards que nous avons élaborés lors de précédents travaux [Aimé *et al.* (2009a),Aimé *et al.* (2010)]. Ces éléments sont, d'une part, des pondérations portées par les liens *isa* dans le treillis de concepts pour rendre compte de la notion de prototypicalité, et, d'autre part, 3 coefficients représentant les poids des composantes sémiotiques intensionnelle, extensionnelle et expressionnelle dans la conceptualisation du domaine.

2.1. Ontologies et prototypicalité

Construire une ontologie O d'un domaine D consiste à spécifier une conceptualisation consensuelle de connaissances. Nous appelons endogroupe l'ensemble des personnes qui partagent la conceptualisation capturée dans l'ontologie. Pour un même domaine, plusieurs ontologies peuvent être définies par différents endogroupes. Nous qualifions ces ontologies d'*Ontologies Vernaculaires de Domaine* (OVD), le terme vernaculaire étant utilisé au sens de relatif à une communauté d'usages, et non au sens de populaire. Nous définissons une *Ontologie Vernaculaire de Domaine* (OVD), pour un domaine D donné et un endogroupe G donné, par le tuple suivant :

$$O_{(D,G)} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\right\} \text{ où}$$

- $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$ sont les ensembles de concepts, de propriétés et d'instances des concepts ;

- $\leq^C$: $\mathcal{C}\times\mathcal{C}$ et $\leq^P$: $\mathcal{P}\times\mathcal{P}$ sont des ordres partiels définissant les hiérarchies de concepts et de propriétés[1] ;
- $dom\ : \mathcal{P} \rightarrow \mathcal{C}$ et $codom\ : \mathcal{P} \rightarrow (\mathcal{C} \cup Datatypes)$ associent à chaque propriété son domaine et éventuellement son co-domaine ;
- $\sigma\ : \mathcal{C} \rightarrow \mathscr{P}(\mathcal{I})$ associe à chaque concept ses instances ;
- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ est le lexique du dialecte de G relatif au domaine D où (1) L_C, L_P et L_I sont les ensembles des termes associés à $\mathcal{C}$, $\mathcal{P}$ et $\mathcal{I}$, et (2) les fonctions $term_c : \mathcal{C} \rightarrow \mathscr{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathscr{P}(L_P)$ et $term_i : \mathcal{I} \rightarrow \mathscr{P}(L_I)$ associent aux primitives conceptuelles les termes qui les désignent.

À ce modèle classique, nous avons proposé dans [Aimé *et al.* (2009a)] et [Aimé *et al.* (2010)] d'ajouter des pondérations permettant de représenter la notion de prototypicalité. Il s'agit de rendre compte dans le modèle du fait que, par exemple, lorsqu'on évoque le concept d'oiseau, l'image qui nous vient à l'esprit est plus proche de celle du moineau que de celle du héron ou du cacatoès. Les sous-concepts du concept d'oiseau doivent donc être ordonnés selon leur représentativité par rapport à leur concept père. Cette *prototypicalité conceptuelle* est représentée dans notre modèle par des pondérations portées par les relations $\leq^C$ entre concepts.

Un autre type de prototypicalité existe entre chaque concept et les termes utilisés pour le désigner. Ainsi, le terme oiseau est préférentiellement utilisé pour désigner ce concept, avant celui de piaf ou de volaille. Cette *prototypicalité lexicale* est représentée dans notre modèle par des pondérations portées par les relations $term_c$ entre concepts et termes. Nous avons proposé dans [Aimé *et al.* (2009a)] et [Aimé *et al.* (2010)] une méthode permettant le calcul semi-automatique de ces deux ensembles de pondérations, nous n'y reviendrons pas ici et considérons qu'elles sont présentes dans l'ontologie.

Nous avons également introduit un vecteur prototype $\overrightarrow{p_{c_p}}$ qui, pour chaque concept c_p, est un vecteur dans l'espace des propriétés $\mathcal{P}$ et représente le prototype du concept en terme intensionnel. La coordonnée de chaque propriété dans le vecteur représente l'importance de la propriété dans la conceptualisation de la notion (cette coordonnée vaut 0 si le concept n'appartient pas au domaine de la propriété). Par exemple, le fait que toutes les voitures ont un

[1] $c_1 \leq^C c_2$ signifie que le concept c_2 subsume le concept c_1.

frein à main n'est pas très important pour ce concept, car ce n'est pas à cette propriété que l'on pense généralement lorsqu'on parle de voiture. Par contre, le fait qu'elles aient toutes des roues est plus important.

Nous avons proposé un mode de calcul des vecteurs prototypes qui suppose que soient disponibles des pondérations qui, pour chaque concept et pour chaque propriété de ce concept, représente l'importance de la propriété pour ce concept. Nous supposerons, pour définir nos mesures, que ces pondérations existent.

2.2. Ontologies sémiotiques

Les modèles ontologiques incorporent généralement les trois dimensions sémiotiques introduites par Morris et Peirce [Morris (1938)] : l'*intension* des concepts, c'est-à-dire la sémantique qui leur est attachée, l'*extension* des concepts, c'est-à-dire leurs instances et l'*expression* des concepts, c'est-à-dire les termes qui les désignent. Mais l'importance de chacune de ces composantes dans la conceptualisation d'un domaine peut varier. Ainsi, les mathématiques sont conceptualisées surtout en intension, à partir des propriétés des concepts. Les catégories d'animaux sont conceptualisées plutôt en extension, à partir des animaux que l'on rencontre (mais un biologiste aura tendance à conceptualiser ces catégories en intension). Le domaine divin, pour un athée ou un agnostique, est conceptualisé en expression, les concepts étant essentiellement des mots sans signification bien définie, et sans instance.

Pour rendre compte dans notre modèle des importances respectives de ces trois composantes intensionnelle, extensionnelle et expressionnelle dans l'ontologie, nous introduisons 3 coefficients respectivement nommés α, β et γ. Nous imposons que ces coefficients soient positifs ou nuls, prennent leurs valeurs dans l'intervalle $[0, 1]$, et que $\alpha+\beta+\gamma = 1$. Les valeurs de ces trois coefficients peuvent être fixées arbitrairement, ou calibrées par expérimentations mais nous avons proposé dans [Aimé *et al.* (2009a, 2010)] une méthode pour en calculer automatiquement des approximations. Nous considérons que le triplet (α, β, γ) caractérise les coordonnées cognitives de l'utilisateur dans le triangle sémiotique.

3. SEMIOSEM, une mesure de similarité sémiotique

SEMIOSEM évalue la similarité entre deux concepts suivant les trois dimensions sémiotiques. Deux concepts sont similaires si (1) d'un point de vue in-

tensionnel, la distance entre leurs vecteurs prototypes est faible, (2) d'un point de vue extensionnel, ils possèdent un grand nombre d'instances en commun et (3) d'un point de vue expressionnel, ils partagent un grand nombre de termes pour les désigner. Nous définissons SEMIOSEM : $\mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ formellement par :

$$SemioSem(c_1, c_2) = \alpha.int_s(c_1, c_2) + \beta.ext_s(c_1, c_2) + \gamma.exp_s(c_1, c_2)$$

Les fonctions int_s, ext_s et exp_s sont respectivement détaillées dans les sections 3.1., 3.2. et 3.3..

3.1. Composante intensionnelle de SEMIOSEM

D'un point de vue *intensionnel*, plus les prototypes respectifs de c_1 et c_2 sont semblables, c'est-à-dire plus les deux concepts partagent de propriétés, plus ces concepts sont similaires, et ceci en tenant compte de l'importance des propriétés dans la définition des deux concepts. La composante intensionnelle $int_s : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est donc calculée comme étant la distance entre les vecteurs prototypes des deux concepts. Cette fonction est définie par :

$$int_s(c_1, c_2) = 1 - dist(\vec{p_{c_1}}, \vec{p_{c_2}})$$

3.2. Composante extensionnelle de SEMIOSEM

D'un point de vue *extensionnel*, nous nous appuyons sur la mesure de similarité de Dice [Dice (1945)]. Cette mesure est définie par le ratio entre le nombre d'instances communes à deux concepts et la moyenne du nombre d'instances des deux concepts. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre d'instances en commun et très peu d'instances distinctes. Cette mesure offre de plus une plus grande régularité que la mesure de Jaccard [Jaccard (1901)]. La fonction $ext_s : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie par :

$$ext_s(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{(|\sigma(c_1)| + |\sigma(c_2)|)/2}$$

3.3. Composante expressionnelle de SEMIOSEM

La fonction exp_s est définie par le ratio entre le nombre de termes communs et le nombre total de termes désignant les deux concepts. Ainsi, deux concepts sont similaires s'ils possèdent un grand nombre de termes communs les désignant et très peu de termes propres à chacun. La fonction exp_s : $\mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est formellement définie par :

$$exp_s(c_1, c_2) = \frac{|term_c(c_1) \cap term_c(c_2)|}{|term_c(c_1) \cup term_c(c_2)|}$$

4. PROXIMA, une mesure de proximité sémiotique

PROXIMA évalue la proximité entre deux concepts suivant les trois dimensions sémiotiques. Cette proximité ne peut seulement être calculée uniquement à partir de l'ontologie, car celle-ci ne contient pas d'éléments permettant d'évaluer la proximité de deux concepts du point de vue expressionnel. La comparaison des termes désignant deux concepts ne permet pas d'apprécier à quel point ces termes sont liés dans l'univers cognitifs de l'endogroupe considéré. Nous utilisons donc un corpus de textes, supposé représentatif de cet univers cognitif, et qui nous sert à calculer les composantes expressionnelle et extensionnelle de PROXIMA.

Sur le principe, deux concepts sont considérés comme proches si (1) d'un point de vue intensionnel, ils sont reliés par de nombreuses propriétés, (2) d'un point de vue extensionnel, les termes désignant leurs instances sont souvent présents ensemble dans le corpus de textes et (3) d'un point de vue expressionnel, les termes qui les désignent sont souvent présents ensemble dans le corpus de textes. Nous définissons PROXIMA : $\mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ formellement par :

$$Proxima(c_1, c_2) = \alpha.int_p(c_1, c_2) + \beta.ext_p(c_1, c_2) + \gamma.exp_p(c_1, c_2)$$

Les fonctions int_p, ext_p et exp_p sont respectivement détaillées dans les sections 4.1., 4.2. et 4.3..

4.1. Composante intensionnelle de PROXIMA

D'un point de vue intensionnel, la proximité entre deux concepts peut être évaluée par le rapport entre le nombre de propriétés qui les lient et le nombre total de propriétés ayant pour domaine l'un des deux concepts. Si au moins une propriété existe entre les deux concepts, la fonction $int_p : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie formellement par :

$$int_p(c_1, c_2) = \left[1 - \log\left(\frac{|p_{12}|+|p_{21}|}{|p_1|+|p_2|}\right)\right]^{-1}$$

$p_1 = \{p_k \in \mathcal{P} : c_1 \in dom(p_k)\}$ est l'ensemble des propriétés ayant c_1 pour domaine

$p_2 = \{p_k \in \mathcal{P} : c_2 \in dom(p_k)\}$ est l'ensemble des propriétés ayant c_2 pour domaine

$p_{12} = \{p_k \in \mathcal{P} : c_1 \in dom(p_k) \wedge c_2 \in codom(p_k)\}$ est l'ensemble des propriétés ayant c_1 pour domaine et c_2 pour co-domaine

$p_{21} = \{p_k \in \mathcal{P} : c_2 \in dom(p_k) \wedge c_1 \in codom(p_k)\}$ est l'ensemble des propriétés ayant c_2 pour domaine et c_1 pour co-domaine

Si $|p_1| + |p_2| = 0$, $int_p(c_1, c_2) = 0$. Nous choisissons de ne pas élargir l'évaluation de la proximité entre concepts au-delà des propriétés qui lient directement les deux concepts. En effet, on constate empiriquement que rien n'assure que deux concepts liés par une chaine de 2 propriétés ou plus soient proches. Par exemple, un article est produit par un auteur, qui a un certain âge, mais les concepts d'article et d'âge ne sont pas proches.

4.2. Composante extensionnelle de PROXIMA

Le calcul de la composante extensionnelle de PROXIMA est similaire à celui de la composante expressionnelle et repose sur l'utilisation d'un corpus de textes représentatifs du domaine pour l'endogroupe considéré. D'un point de vue extensionnel, nous considérons que deux concepts sont d'autant plus proches que les termes désignant leurs instances sont présents ensemble dans les mêmes documents. La fonction $ext_p : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est formellement définie par :

$$ext_p(c_1, c_2) = \frac{nbBoth_i(c_1,c_2)}{nbOne_i(c_1,c_2)}$$

$nbBoth_i(c_1, c_2)$ est le nombre de documents où au moins un terme désignant une instance de c_1 et au moins un terme désignant une instance de c_2 sont présents ensemble, et $nbOne_i(c_1, c_2)$ est le nombre de documents où au moins un terme désignant une instance de c_1 ou de c_2 est présent.

Une matrice $TFIDF_i$ est calculée dans $[0, 1]^{t \times d}$, où t est le nombre de termes apparaissant dans l'ontologie pour désigner les instances et d le nombre de documents du corpus. On a :

$$nbBoth_i(c_1, c_2) = \sum_{d_j \in DBoth_i(c_1,c_2)} \sum_{t_i \in T_i(c_1,c_2)} TFIDF_i(i, j)$$

Avec $DBoth_i(c_1, c_2) = \{d_z : (term_i(c_1) \cap d_z \neq \emptyset) \wedge (term_i(c_2) \cap d_z \neq \emptyset)\}$ et $T_i(c_1, c_2) = term_i(c_1) \cup term_i(c_2)$.

$$nbOne_i(c_1, c_2) = \sum_{d_j \in DOne_i(c_1,c_2)} \sum_{t_i \in T_i(c_1,c_2)} TFIDF_i(i, j)$$

Avec $DOne_i(c_1, c_2) = \{d_z : (term_i(c_1) \cap d_z \neq \emptyset) \vee (term_i(c_2) \cap d_z \neq \emptyset)\}$ et $T_i(c_1, c_2) = term_i(c_1) \cup term_i(c_2)$.

4.3. Composante expressionnelle de PROXIMA

La composante expressionnelle est calculée à partir d'un corpus de textes représentatifs du domaine pour l'endogroupe considéré. D'un point de vue expressionnel, plus deux concepts sont désignés par des termes présents ensemble dans les mêmes documents, plus ils sont proches. Plus précisément, la composante expressionnelle de PROXIMA est d'autant plus élevée que le nombre de documents regroupant à la fois des termes désignant les deux concepts est élevé, et ceci en comparaison du nombre de documents où apparait au moins un des deux concepts. La fonction $exp_p : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie formellement par :

$$exp_p(c_1, c_2) = \frac{nbBoth_c(c_1,c_2)}{nbOne_c(c_1,c_2)}$$

$nbBoth_c(c_1, c_2)$ est le nombre de documents où au moins un terme désignant le concept c_1 et au moins un terme désignant le concept c_2 sont présents, et $nbOne_c(c_1, c_2)$ est le nombre de documents où au moins un terme désignant c_1 ou c_2 est présent.

Le simple comptage du nombre de documents où les termes apparaissent ne rend cependant pas compte de l'importance de ces termes dans les documents. Une des meilleures mesures de cette importance est TF-IDF (Term Frequency - Inverse Document Frequency) [Salton & McGill (1986)]. Pour un terme t_i et un document d_j, le TF-IDF est donné par :

$$TFIDF(t_i, d_j) = \frac{n_{i,j}}{\sum_k n_{k,j}} . log \frac{|D|}{|\{d_k : t_i \in d_k\}|}$$

$n_{i,j}$ est le nombre d'occurrences du terme t_i dans le document d_j

$\sum_k n_{k,j}$ est le nombre d'occurrences de tous les termes dans le document d_j

$|D|$ est le nombre de documents du corpus et $|\{d_k : t_i \in d_k\}|$ est le nombre de documents où le terme t_i apparaît

Nous proposons donc pour calculer $nbBoth_c(c_1, c_2)$ de sommer les valeurs de TF-IDF de chaque terme désignant c_1 ou c_2 apparaissant dans les documents où sont présents à la fois des termes désignant c_1 et des termes désignant c_2.

En pratique, nous calculons une matrice $TFIDF_c$ dans $[0, 1]^{t \times d}$, où t est le nombre de termes apparaissant dans l'ontologie pour désigner les concepts et d le nombre de documents du corpus. $TFIDF_c(i, j)$ vaut 0 si le terme t_i n'apparait pas dans le document d_j et vaut $TFIDF(t_i, d_j)$ sinon.

Le calcul de $nbBoth_c(c_1, c_2)$ consiste alors à sommer les coefficients de la matrice $TFIDF_c$ pour tous les termes désignant c_1 ou désignant c_2 et pour tous les documents contenant à la fois un terme désignant c_1 et un terme désignant c_2 :

$$nbBoth_c(c_1, c_2) = \sum_{d_j \in DBoth_c(c_1,c_2)} \sum_{t_i \in T_c(c_1,c_2)} TFIDF_c(i, j)$$

Avec $DBoth_c(c_1, c_2) = \{d_z : (term_c(c_1) \cap d_z \neq \emptyset) \wedge (term_c(c_2) \cap d_z \neq \emptyset)\}$ et $T_c(c_1, c_2) = term_c(c_1) \cup term_c(c_2)$.

Le terme $nbOne_c(c_1, c_2)$ se calcule de la même manière à partir de la matrice $TFIDF_c$ mais en sommant les coefficients pour tous les termes désignant c_1 ou désignant c_2 et pour tous les documents contenant un terme désignant c_1 ou un terme désignant c_2 :

$$nbOne_c(c_1, c_2) = \sum_{d_j \in DOne_c(c_1,c_2)} \sum_{t_i \in T_c(c_1,c_2)} TFIDF_c(i, j)$$

Avec $DOne_c(c_1, c_2) = \{d_z : (term_c(c_1) \cap d_z \neq \emptyset) \vee (term_c(c_2) \cap d_z \neq \emptyset)\}$ et $T_c(c_1, c_2) = term_c(c_1) \cup term_c(c_2)$.

La composante expressionnelle donnée ici mesure bien une proximité entre concepts, relativement à un corpus de documents, mais ne donne pas d'information sur la similarité entre les deux concepts. Mais deux concepts peuvent bien entendu être proches dans un corpus et similaires. C'est le cas par exemple des concepts de conclusion et d'introduction, qui apparaissent souvent dans les mêmes documents, et sont assez similaires (modulo des fonctionnalités différentes). Deux concepts peuvent également être proches dans

un corpus et non similaires. Par exemple les termes bouteille et vin se trouvent souvent ensemble, mais ne désignent pas des objets similaires.

5. Expérimentations

Tester toutes les composantes de SEMIOSEM et PROXIMA requiert à la fois une ontologie relativement riche (avec en particulier des propriétés entre concepts et des instances pour les différents concepts) mais également des pondérations représentant l'importance de chaque propriété dans la définition de chaque concept (pour le calcul de la composante intensionnelle de SEMIOSEM) et un corpus de textes représentatifs du domaine (pour le calcul des composantes extensionnelle et expressionnelle de PROXIMA).

Ne disposant pas d'ontologie comportant de nombreuses instances, nous n'avons pas tenu compte dans les tests des composantes extensionnelles. De plus, SEMIOSEM n'a été calculée qu'à partir de sa composante expressionnelle. Deux tests ont été menés : le premier a permis de comparer nos mesures aux mesures de similarité existantes ; le deuxième a permis d'évaluer les deux mesures sur un jeu d'essai restreint pour s'assurer que leur comportement est en adéquation avec les notions de proximité et de similarité que nous proposons.

5.1. Comparaison avec les mesures existantes

Nous avons retenu pour ce test les mesures de similarité proposées par Jiang et Conrath [Jiang & Conrath (1997)], Wu et Palmer [Wu & Palmer (1994)], Resnik [Resnik (1999)], Hirst et St-Onge [Hirst & St-Onge (1998)] et Lin [Lin (1998)]. Pour disposer d'un jeu d'essai conséquent, nous avons choisi d'utiliser l'ontologie WordNet 3.0 et le corpus Corpus of Contemporary American English[2] issu de magazines, journaux, romans et textes académiques (de 1920 à 2010) et comportant environ 160000 textes et près de 410 millions de termes.

À partir de ces ressources, et en utilisant l'outil Wordnet Similarity de S. Patwardhan et T. Pedersen[3][Patwardhan *et al.* (2003)], nous avons calculé les similarités (et proximités) pour les 353 couples de mots fournis par la version de 2004 du test WordSimilarity-353[4] de [Finkelstein *et al.* (2002)]. Ce

[2]http://www.americancorpus.org/

[3]http://www.d.umn.edu/tpederse/similarity.html

[4]http://www.cs.technion.ac.il/gabr/resources/data/wordsim353/wordsim353.html

test fournit également pour chaque couple une valeur de similarité fixée par une dizaine de personnes et nous avons calculé le coefficient de corrélation linéaire entre les valeurs calculées et les valeurs fournies par WordSimilarity. Seule la composante expressionnelle de SEMIOSEM et les composantes expressionnelle et intensionnelle de PROXIMA sont calculées ici. Pour le calcul de PROXIMA, nous avons fixé les coordonnées sémiotiques à $\alpha = 0.66, \beta = 0, \gamma = 0.33$ (et $\alpha = 0, \beta = 0, \gamma = 1$ pour SEMIOSEM).

Mesure	*Corrélation*
Jiang & Conrath	0.61
Lin	0.7
Resnik	0.78
Wu & Palmer	0.75
Hirst & St Onge	0.64
PROXIMA	0.45
SEMIOSEM	0.71

Table 1: Coefficients de corrélation linéaire WordSimilarity-353 / mesures.

Le tableau 1 regroupe les résultats obtenus. SEMIOSEM présente un coefficient de corrélation satisfaisant comparé aux autres mesures alors que PROXIMA obtient un score faible. Mais si on regarde les résultats en détails, on s'aperçoit que les mesures existantes, tout comme les valeurs fournies par WordSimilarity, agrègent similarité et proximité. Par exemple, les termes *cup* et *coffee* obtiennent un score de 6.58 dans le test WordSimilarity, et des valeurs élevées avec les mesures existantes, alors que le calcul de SEMIOSEM sur cet exemple donne une valeur quasi nulle, ce qui est davantage conforme à la sémantique de ces concepts.

Pour évaluer plus précisément SEMIOSEM et PROXIMA, nous avons donc construit deux tests séparant autant que possible les notions de proximité et de similarité.

5.2. Comparaison entre SEMIOSEM et PROXIMA

Nous avons comparé nos mesures avec un jugement humain de proximité et un jugement humain de similarité. Pour ce faire, nous avons extrait 31 couples de mots du test WordSimilarity-353 : (1) un groupe de 10 couples de concepts estimés peu ou pas similaires, avec une note faible, (2) un groupe

de 10 couples de concepts estimés très similaires, avec une note élevée, et (3) un groupe de 11 couples de concepts estimés peu similaires, avec une note intermédiaire. La liste de couples, classée aléatoirement, a été soumise à un ensemble de 22 personnes auxquelles nous avons demandé de donner une note de proximité et une note de similarité sur la même échelle que le test WordSimilarity-353. Les moyennes des notes ont été retenues pour le test. Les tableaux 2 et 3 regroupent des exemples de valeurs de similarité et proximité obtenues.

Mot 1	*Mot 2*	*Similarité*	SEMIOSEM
pierre précieuse	joyau	*9.82*	*1.00*
rue	avenue	*8.88*	*0.75*
tasse	café	*5*	*0.00*

Table 2: Extrait des valeurs obtenues avec SEMIOSEM.

Sur ce jeu d'essai réduit, nous obtenons un coefficient de corrélation linéaire de 0.69 entre les valeurs données par SEMIOSEM et celles du jugement humain de similarité. SEMIOSEM obtient donc un score un peu plus faible que sur le test WordSimilarity. Mais le détail des valeurs du jugement humain montre encore une fois que la proximité entre concepts influe sur l'évaluation humaine de la similarité. Par exemple, le couple de mots *tasse* et *café* obtient une similarité de 5 dans le jugement humain de similarité, ce qui est conceptuellement erroné.

D'autre part, il nous parait que la place de l'intension des concepts est prépondérante dans le processus de jugement de similarité. En d'autres termes, on estime la similarité en se basant avant tout sur les propriétés des concepts. Le calcul de SEMIOSEM sur la seule composante expressionnelle ne peut donc pas donner un très bon résultat.

Mot 1	*Mot 2*	*Proximité*	PROXEM
pierre précieuse	joyau	*9.6364*	*0.50730*
rue	avenue	*7.9091*	*0.45517*
tasse	café	*6.5455*	*0.41793*

Table 3: Extrait des valeurs obtenues avec PROXIMA.

Nous obtenons un coefficient de corrélation linéaire de 0.81 entre les valeurs données par PROXIMA et celles du jugement humain de proximité. PROXIMA obtient donc un très bon score lorsque le distinguo entre proximité et similarité est fait. D'autre part, ces valeurs sont obtenues avec un poids de 0.66 pour la composante intensionnelle, ce qui semble indiquer que la proximité entre concepts est fortement corrélée à l'existence de propriétés liant ces concepts.

6. Conclusion et perspectives

Nous avons proposé deux nouvelles mesures de similarité et proximité entre concepts, qui ont pour originalité d'une part de bien distinguer ces deux notions et d'autre part de tirer parti des trois composantes sémiotiques des concepts. SEMIOSEM évalue la similarité entre deux concepts, c'est-à-dire leur analogie, en s'appuyant sur la comparaison entre leurs prototypes et sur le fait qu'ils partagent des instances communes et des termes communs. PROXIMA évalue la proximité entre deux concepts, c'est-à-dire le fait qu'ils soient cognitivement liés, en mesurant la densité des propriétés qui les relient, et l'apparition conjointe dans un corpus de textes des termes désignant ces concepts ou leurs instances.

Nos mesures peuvent également s'adapter au domaine de connaissances et aux utilisateurs via la pondération des composantes sémiotiques, pondérations qui représentent l'importance respective des aspects intensionnel, extensionnel et expressionnel dans la conceptualisation du domaine. Cette richesse et cette adaptabilité de nos mesures n'est cependant pleinement exploitable que si certaines ressources sont disponibles pour les calculer, en particulier des instances en nombre suffisant pour la composante extensionnelle.

Par rapport aux mesures utilisant les liens hiérarchiques, un des avantages de SEMIOSEM et PROXIMA est d'être indépendantes de la structure hiérarchique de l'ontologie. En effet, les valeurs de similarité et de proximité ne changeront pas si on introduit dans la hiérarchie des concepts additionnels pour structurer davantage l'ontologie, ou si on retire des concepts. Il nous semble qu'il s'agit là d'un point important, du fait que les ontologies sont de plus en plus structurées verticalement selon un schéma ontologie fondationnelle/ontologie noyau/ontologie de domaine, et incorporent donc de plus en plus de concepts servant de ciment à l'ontologie, mais sans signification réelle pour les utilisateurs, et pouvant donc fausser les calculs de similarité basés sur la structure.

Cependant, l'utilisation intensive de nos mesures pose des problèmes techniques. Tout d'abord, pour le calcul de la composante intensionnelle de SEMIOSEM, la pondération des propriétés peut s'avérer impraticable pour des ontologies de très grande taille. Il est possible de mettre tous ces poids à 1 pour calculer la composante, mais elle perd alors de son intérêt. D'autre part, le temps de calcul du nombre d'occurrences de termes dans les textes devient conséquent si le corpus est de très grande taille (cependant, ce calcul ne se fait qu'une seule fois). Les résultats obtenus sont par ailleurs nettement dépendants de la qualité du corpus. Enfin, SEMIOSEM est dépendante de l'imprécision des calculs d'occurrences liée aux limites du TALN.

Les premiers tests montrent que nos mesures se révèlent pertinentes et s'avèrent convenablement corrélées au jugement humain. Cependant, des expériences plus larges, et portant sur toutes les composantes de nos mesures, restent à mener pour réellement valider notre approche. D'autre part, élargir les tests permettra d'étudier plus précisément les liens entre similarité, proximité et représentation ontologique des connaissances. Une expérience intéressante sera par exemple de faire varier les poids des composantes pour estimer les influences respectives des aspects intensionnel, extensionnel et expressionnel dans l'évaluation humaine des similarités et proximités.

References

AIMÉ X., FÜRST F., KUNTZ P. & TRICHET F. (2009a). Gradients de prototypicalité appliqués à la personnalisation d'ontologies. In *IC 2009 : Actes des 20es Journées Francophones d'Ingénierie des Connaissances IC 2009*, p. 241–252: PUG.

AIMÉ X., FÜRST F., KUNTZ P. & TRICHET F. (2009b). Semiosem : une mesure de similarité conceptuelle fondée sur une approche sémiotique. In *Actes d'IC*, p. 229–240: PUG.

AIMÉ X., FURST F., KUNTZ P. & TRICHET F. (2010). Prototypicality gradient and similarity measure: a semiotic-based approach dedicated to ontology personalization. *Journal of Intelligent Information Management. Scientific Research*, **2**(2), 65–79.

BUDANITSKY A. & HIRST G. (2006). Evaluating word-based measures of semantic distance. *Computationnal Linguistics*, **32**(1), 13–37.

DICE L. (1945). Measures of the amount of ecological association between species. *Ecology*, **26**, 297–302.

EUZENAT J. & VALTCHEV P. (2004). Similarity-based ontology alignment in OWL-Lite. In R. L. DE MANTARAS & L. SAITTA, Eds., *European Conference on Artificial Intelligence (ECAI'2004)*, p. 333–337: IOS Press.

FINKELSTEIN L., GABRILOVICH E., MATIAS Y., RIVLIN E., SOLAN Z., WOLFMAN G. & RUPPIN E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

HIRST G. & ST-ONGE D. (1998). Lexical chains as representation of context for the detection and correction malapropisms. *WordNet: An electronic lexical database and some of its applications*, p. 305–332.

HLIAOUTAKIS A., VARELAS G., VOUTSAKIS E., PETRAKIS E. G. M. & MILIOS E. (2006). Information retrieval by semantic similarity. **3**(3), 55–73.

JACCARD P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise de Sciences Naturelles*, **37**, 241–272.

JIANG J. & CONRATH D. (1997). Semantic similarity based on corpus statistics and lexical taxinomy. In *International Conference en Research in Computationnal Linguistics*, p. 19–33.

KOFFKA K. (1935). *Principles of Gestalt Psychology*. Routledge & Kegan Paul PLC.

LIN D. (1998). An information-theoric definition of similarity. In *Proceedings of the 15th international conference on Machine Learning*, p. 296–304.

MORRIS C. (1938). *Foundations of the Theory of Signs*. Chicago University Press.

PATWARDHAN S., BANERJEE S. & PEDERSEN T. (2003). Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, p. 241–257.

QUILLIAN M. (1968). Semantic memory. In *Semantic Information Processing*: MIT Press.

RESNIK P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, **11**, 95–130.

SALTON G. & MCGILL M. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill.

WU Z. & PALMER M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, p. 133–138.